Interactive comment on "On the importance of multiple-component evaluation of spatial patterns for optimization of earth system models – A case study using mHM v5.6 at catchment scale" by Julian Koch et al.

Anonymous Referee #1
Received and published: 29 December 2017

1 Summary
The paper presents new metric that evaluates the spatial pattern of hydrologic model and earth system model. The new metric called SPAEF is multi-objectives, and consists of three components; spatial correlation, coefficient of variance ratio (simulation to observation), and histogram matching. The paper demonstrated mHM hydrologic model calibration by applying this metric to simulated ET distribution (or latent heat flux) against remote sensing data over 2500 sq-km catchment in Denmark and compared the calibration performance against the use of the other metrics. The paper show that updated parameterization improves ET spatial pattern over use of the previous model parameters.

We would like to thank the reviewer for his/her thorough revision of our manuscript. We are very pleased that our work on spatial pattern oriented model evaluation is generally well received by the reviewer. The comments raised by the reviewer pose valuable thoughts and the rigorous revision following his/her suggestions will certainly improve the scientific quality of our work. Our replies below indicate what we intend to change in the manuscript prior to resubmission.

2 Comments
Goals of this paper, which is to propose new evaluation/calibration metric that quantifies the accuracy of spatial pattern of the earth system model, is good fit for GMD. Overall, I, as hydrologists who do modeling work, enjoyed reading the manuscript with great interest. My main comments below are regarding how this metrics and calibration strategy could be applied to the other model than mHMs, which might be hard to estimate spatially distributed parameters. My recommendation would be minor revision (if you can justify not performing additional simulations I mention in comment 4

1. To promote the metrics invented here, acronym of the metric is better pronounceable. Also, I would consider the metric name in Title. Just suggestion.

We will follow the reviewer's advice and add the name of the metric including acronym in the title: "The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models". We agree that "SPAEF" may not be easy to pronounce, but this is nothing we have considered during the formulation of the metric. Also other popular metrics such as KGE or NSE are also not easily pronounceable

2. Please describe the weakness of two other metrics you evaluated besides SPAEF clearly.

We will add a clear discussion of the differences between SPAEF and Connectivity and FSS in the revised manuscript. Figure 4 as well as Table 1 can be used as illustrations to elaborate on the differences between the metrics. In comparison to SPAEF, Connectivity does not consider variability or the correct allocation. FSS constrains the right allocation but also does not explicitly handle variability. However, it may not be a completely fair comparison, because we argue that multiple components have to be taken into consideration when comparing spatial patterns. FSS and Connectivity have their strengths, but are single component metrics which perform less satisfactory in comparison to SPAEF. SPAEF is a multiple component metric which marks the key advantage over the other two metrics.

3. The paper stated that spatial pattern of the model outputs depends at least on 1) process parameterizations (i.e., model equations), 2) accuracy of climate forcing (spatio-temporal pattern), and 3) parameter regionalization scheme (how parameters are distributed in space). I agree with these, but I speculate that spatial pattern is regulated in the first order by transfer function forms that convert soil/vegetation data to parameter values. Maybe mention this?

We agree with the reviewer on this point, the transfer functions were the key element that allowed us to obtain such a satisfying result in terms of spatial pattern performance. However, the remaining two points are still relevant. The catchment used for this study is characterized by quite homogeneous climatic forcing and the monthly maps of ET are therefore less effected by climate in comparison to soil and vegetation. The spatial pattern calibration of a catchment with a strong climate gradient may be more constrained by the quality of the climate forcing than the Skjern catchment. Lastly, having the right process descriptions is essential to predict any physical system. We will make sure to point out the importance of the transfer functions in the revised manuscript.

4. While mHM has a very unique regionalization scheme called mulit-scale parameter regionalization scheme (calibrate the coefficients of transfer functions that compute parameter values from distributed geophysical data), making it easy to regionalize the parameters at any scales, all most all the other models do not have such a scheme. Therefore, it seems to be difficult to perform distributed model calibration presented in this paper for the other models. How applicable is this calibration strategy to the other models?

This is right, MPR allows easy regionalization in mHM, but its application is not limited to mHM. MPR can also be added to other model structures, as presented by Samaniego et al. (2017) for PCR-GLOBWB and Mizukami et al. (2017) *for VIC. Samaniego et al. (2017) have outlined a modelling protocol to describe how MPR can be added to a particular model, which extends the applicability of MPR beyond mHM. We will provide the two references below and a discussion of the transferability of MPR to other models in the revised manuscript. Besides MPR, which is one way to implement parameter regionalization, in the calibration of distributed models, every modeler should think of way to regionalize parameters during calibration. This can be by self-implemented transfer functions which are added as a pre-processing script to the calibration routine. Regionalization is certainly not limited to MPR and simpler solutions may be sufficient in some cases to give the parameter fields the desired freedom to adjust a simulated to an observed spatial pattern.*

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O. and Samaniego, L.: Towards seamless large-domain parameter estimation for

hydrologic models, Water Resour. Res., 53(9), 8020–8040, doi:10.1002/2017WR020401, 2017.
Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E., Warrach-Sagi, K. and Attinger, S.: Toward seamless hydrologic predictions across spatial scales, Hydrol. Earth Syst. Sci., 21(9), 4323–4346, doi:10.5194/hess-21-4323-2017, 2017.

5. However, I still think this is an unique calibration strategy that combines spatial pattern and temporal pattern metrics, but meantime, I thought there need for more calibration experiments to understand the values of spatial pattern metrics for calibration purpose. I wish that there would have been results from 1) stream-flow only calibration and 2) spatial pattern metric only calibration, showing skills of both ET spatial pattern and streamflow simulation. This way, the paper could show real value of this spatial pattern calibration. Does streamflow only calibration produce worse ET spatial pattern than the streamflow and ET combined calibration? Does spatial pattern only calibration produce worse streamflow simulations than the case streamflow is not used for calibration?

The reviewer touches upon a very interesting point. Here we would like to refer to Demirel et al. (2017) *who have conducted the above mention calibration experiments for the same model setup. They tested three calibration strategies: A calibration ensemble of Q-only, Spatial-only and a combination of Q and Spatial. Their findings underline the strength of combining temporal and spatial observations, as the uncertainty of predicting Q for the combined calibration was lower than the Q-only calibration. On the other hand, it was not possible for the Spatial-only calibration to constrain the hydrograph in a meaningful way. With respect to the spatial pattern performance, the Q-only calibration resulted in poor spatial patterns while very limited tradeoffs were noticeable comparing the spatial pattern performance of Spatial-only and the combined calibration. This underlines the limited trade-off between Q dynamics and spatial patterns illustrating the benefit of combining observation types in a multi-objective framework. We will refer to the results by Demirel et al. (2017) in the revised manuscript in detail to make the reader aware of the limited tradeoffs between temporal and spatial observations and the fact that spatial patterns have the power to constrain the hydrograph simulation efficiently when being paired with Q observations in a multi-objective calibration framework.*

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L. and Stisen, S.: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, Hydrol. Earth Syst. Sci. Discuss., 1–22, doi:10.5194/hess-2017-570, 2017.

6. Contrast to hydrologic models, earth system model community do not have calibrate the parameters though Land surface model community started to pay more attention to calibrations/sensitivity analysis. Therefore, the presentation of this paper is more related to hydrologic model application. However, spatial pattern metrics could be used for model evaluation purpose. For example, would it be possible (or worthwhile) to use this for evaluation of meteorological fields from climate models against observation or reanalysis grid.

We completely agree to this point which has also been pointed out by reviewer 2. We decided to remove the emphasize on earth system models in the title and introduction and rather focus on the applicability of SPAEF for hydrological models. We will follow the suggestion of the *reviewer and add references which promote the usability of spatial pattern metrics to evaluate spatial patterns of metrological or atmospherical models*

3 Minor comments or specific line by line comments
• I found a few typos – mayor-> major (P2, Line 2), patter->pattern (P5, Line 20).

Thanks, these will be corrected.

• P5, Line3-4.I am not sure if I understand this sentence. Do you mean soil/vegetation properties by "these"?

Exactly, we will change the sentence and try to be more specific.

• P5. Q in KGE equation is incorrect. It should be μ_{sim}/μ_{obs}. Also, correct explanation in Line 14.

Correct. We will update the bias terms in equation 3.

• P6, Line1-9. I think this paragraph is better fit after P5, L18.

Agree. We will reorder this section.

• P9, Line6-7. Use of spatial pattern metrics as objective function converge faster than streamflow derived objective function. That seems to make sense be- cause spatial pattern is by large determined by fixed transfer function forms and soil/vegetation properties in the mHM. It would be nice to mention the reason if you know.

We actually do not compare convergence rates between spatial and temporal objective functions, because we do not show any results that could support such a conclusion. Based on our results we comment on the convergence of the spatial objective functions which support our number of maximum runs for the calibration.

• P10, Line10-14. I think this is good points to discuss, but I think it would be nice to discuss constrains from transfer function form (regularization equations).

We will add a few points on the limitations of MPR, such as that the selection and definition of robust transfer functions can be difficult and bears uncertainties. Reliable transfer functions are crucial for the applicability of MPR. Other limitations are that the transfer functions are tedious to implement in other models besides mHM, as discussed above. Also, the minimum scale at which a model can be applied is depending on the data availability, since the subgrid variability is fundamental to MPR. The abovementioned limitations, among others, are discussed by Samaniego et al. (2017).

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E., Warrach-Sagi, K. and Attinger, S.: Toward seamless hydrologic predictions across spatial scales, Hydrol. Earth Syst. Sci., 21(9), 4323–4346, doi:10.5194/hess-21-4323-2017, 2017.

• P11. Line 22. This number of iteration for convergences should depend on model choice and also regionalization scheme. So it is better not to generalize the conclusion here (I think).

Yes, we will down tone this conclusion and clearly state that this may only be relevant for our study.

• P11. Line26. I don't understand why it is reasonable given the parameterization of the mHM? Please elaborate a little more.

The relationship between histo match and correlation seems reasonable because of the slightly skewed distribution of the ET pattern (Figure 3). The lower side of the distribution are the forest grids, which have a lower ET during the growing season than the agricultural areas. Calibrating against histo match with such a peculiar distribution will result in a reasonable correlation, because low and high values will automatically be allocated correctly. This finding does not result in a crucial conclusion and it is further very much limited to this study and to the applied reference pattern. Therefore we will consider omitting these sentences in the revised version of the manuscript.